Published online 2 June 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W185–W188***
doi:10.1093/nar/gkp434

OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences

Fabian Schreiber[1,2,*], Gert Wörheide[2] and Burkhard Morgenstern[1]

[1]Institut für Mikrobiologie und Genetik, Abteilung für Bioinformatik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077, Göttingen, and [2]Molecular Geo- & Palaeobiology, Department of Earth- and Environmental Sciences & GeoBio-Center LMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Straße 10, 80333 München, Germany

Received February 13, 2009; Revised April 24, 2009; Accepted May 11, 2009

ABSTRACT

In the absence of whole genome sequences for many organisms, the use of expressed sequence tags (EST) offers an affordable approach for researchers conducting phylogenetic analyses to gain insight about the evolutionary history of organisms. Reliable alignments for phylogenomic analyses are based on orthologous gene sequences from different taxa. So far, researchers have not sufficiently tackled the problem of the completely automated construction of such datasets. Existing software tools are either semi-automated, covering only part of the necessary data processing, or implemented as a pipeline, requiring the installation and configuration of a cascade of external tools, which may be time-consuming and hard to manage. To simplify data set construction for phylogenomic studies, we set up a web server that uses our recently developed OrthoSelect approach. To the best of our knowledge, our web server is the first web-based EST analysis pipeline that allows the detection of orthologous gene sequences in EST libraries and outputs orthologous gene alignments. Additionally, OrthoSelect provides the user with an extensive results section that lists and visualizes all important results, such as annotations, data matrices for each gene/taxon and orthologous gene alignments. The web server is available at http://orthoselect.gobics.de.

INTRODUCTION

The rapid development of genome-sequencing techniques has led to the generation of complete genome sequences for >600 species. Most of these sequences belong to model organisms, covering only small portions of the tree of life. The generation of massive numbers of expressed sequence tag (EST) libraries that can now be sequenced inexpensively by third-generation sequencing is a cheap alternative to whole genome sequencing, and has also provided a wealth of phylogenetically relevant data. Several recent phylogenomic studies have used EST sequences to generate large data matrices (1–4).

These studies generated and assembled EST sequences, which were screened for orthologous sequence regions to build useful orthologous gene alignments. Orthologous sequences result from a speciation event, and are likely to have a conserved function, whereas paralogous sequences evolve through a gene duplication event within a species, and are less likely to maintain their original function, due to processes such as neo-/or sub-functionalization (5).

Orthologous and paralogous together are called homologues (6). Since the prime goal of building reliable phylogenetic trees is to decipher the evolutionary relationships among organisms based on their shared common ancestry, only orthologous sequences should be used.

A reliable protocol is needed to build sets of orthologous sequences from EST libraries for successive phylogenomic analyses. We recently proposed such a protocol, which we called OrthoSelect (Schreiber *et al.*, manuscript submitted). The workflow of the protocol is outlined in Figure 1. The main idea is to keep user interaction simple, by simultaneously using state-of-the-art methods for orthology assignment, EST translation and elimination of paralogues, as well as construction and automated refinement of multiple sequence alignments. OrthoSelect has been extensively tested and proven to be a useful tool for managing this complex task.

Here, we present a web interface to OrthoSelect, the first web-based EST analysis pipeline for constructing orthologous gene alignments from EST libraries. Our web

*To whom correspondence should be addressed. Tel: +49 (0) 551 3913884; Fax: +49 551 3914929; Email: Fschrei@gwdg.de

server does not require any kind of installation or testing. The user simply uploads EST libraries and chooses parameters (or uses default settings) to conduct the analysis. OrthoSelect then provides the user with a dataset useful for subsequent phylogenetic analysis, as well as numerous helpful data and statistics, such as annotations, a data matrix showing EST assignments to the orthologous groups (OGs) and visualizations of the orthologous gene alignments.

WEB SERVER

The main purpose of our web server is the construction of orthologous gene alignments from assembled EST libraries or other nucleotide sequences. After the user uploads pools of EST sequences, ESTs are assigned to OGs and the sequences most likely to be orthologous—in case there were multiple sequences per species—are used to compute an alignment that is post-processed in a final step. The workflow of the pipeline is depicted in Figure 1, and is described in more detail in the next section.

METHODS

Using the OGs defined by the eukaryotic orthologous groups (KOG) database (7), each EST is assigned to the closest OG. The assignment is done using a reimplementation of BLASTO (8) that clusters hits from a similarity search of the EST against the KOG database. The similarity between a query sequence and an OG is defined as the mean *E*-value between the query and the sequences from the OG. ESTs are then translated using a standard six-frame translation method.

We then translate the ESTs using the tools ESTScan (9) and GeneWise, (10) to account for frame shift errors. Considering the best Blast hit of the EST as a reference sequence, our program selects the translated sequence that is most similar to the reference sequence. Only OGs with at least three taxa are further considered.

At this stage of the analysis, it is possible to preselect individual or groups of taxa. The set of OGs is then further reduced to contain only OGs containing all preselected taxa. Redundant (e.g. paralogous) sequences are removed from each OG. This is done by considering only the sequence from each species that maximizes a global alignment score as being most likely orthologous. All sequences from each orthologous group are then aligned using either Muscle (11), T-Coffee (12) or DIALIGN-TX (13). These alignments are used to build hidden Markov models (14) that will be used to search the EST libraries for additional hits. Gblocks (15) is subsequently used to remove ambiguously aligned alignment columns. Since EST sequences may only partially cover genes, there is an option to exclude sequences from the alignment that are too short. This procedure outputs gene alignments whose member sequences are the ones most likely to be orthologous, given the dataset.

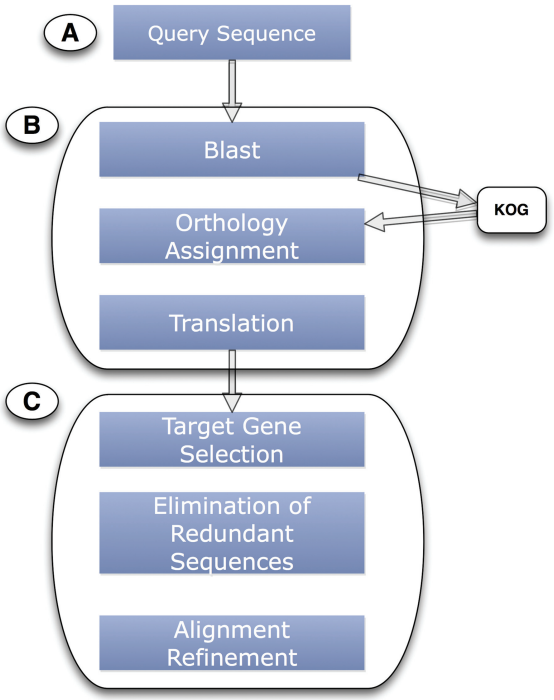

Figure 1. The main workflow of OrthoSelect: Each EST (**A**) is assigned to a pre-defined orthologous group (OG) by the KOG database, and translated (**B**). After all ESTs have been assigned to OGs, a subset of the OGs can be selected which will be further processed to exclude all redundant sequences, compute a sequence alignment and refine it in the last step (**C**).

INPUT

Our web server allows the use of OrthoSelect with default or adapted parameter values, e.g. the *E*-value for similarity searches using Blast (16) or methods for computing multiple sequence alignments. OrthoSelect accepts nucleotide sequences in FASTA format.

In the absence of a standard format for sequence identifiers in FASTA headers, sequence identifiers have to be adapted at some stage of a phylogenetic analysis to allow viewing taxa.

OrthoSelect requires the FASTA header to be in a certain format (the first word up to the first whitespace is taken as an accession number), and uploaded files have to match that format, or can be adapted using a converter supplied on the web page. Several syntax checks for the uploaded EST sequences have been implemented to ensure optimal performance from OrthoSelect. Our web interface offers the possibility to set up sequence identifiers (e.g. abbreviated taxon names) that will be used throughout the analysis.

Furthermore, the user can preselect one taxon or several taxa.

Our web server will then return a list of those OGs to which the submitted ESTs have been assigned, as well as a subset of those OGs containing the preselected taxa. The maximum number of input sequences is 30 000, and the maximum number of EST libraries is 10. An email address has to be supplied, since notification about the results will be sent via email.

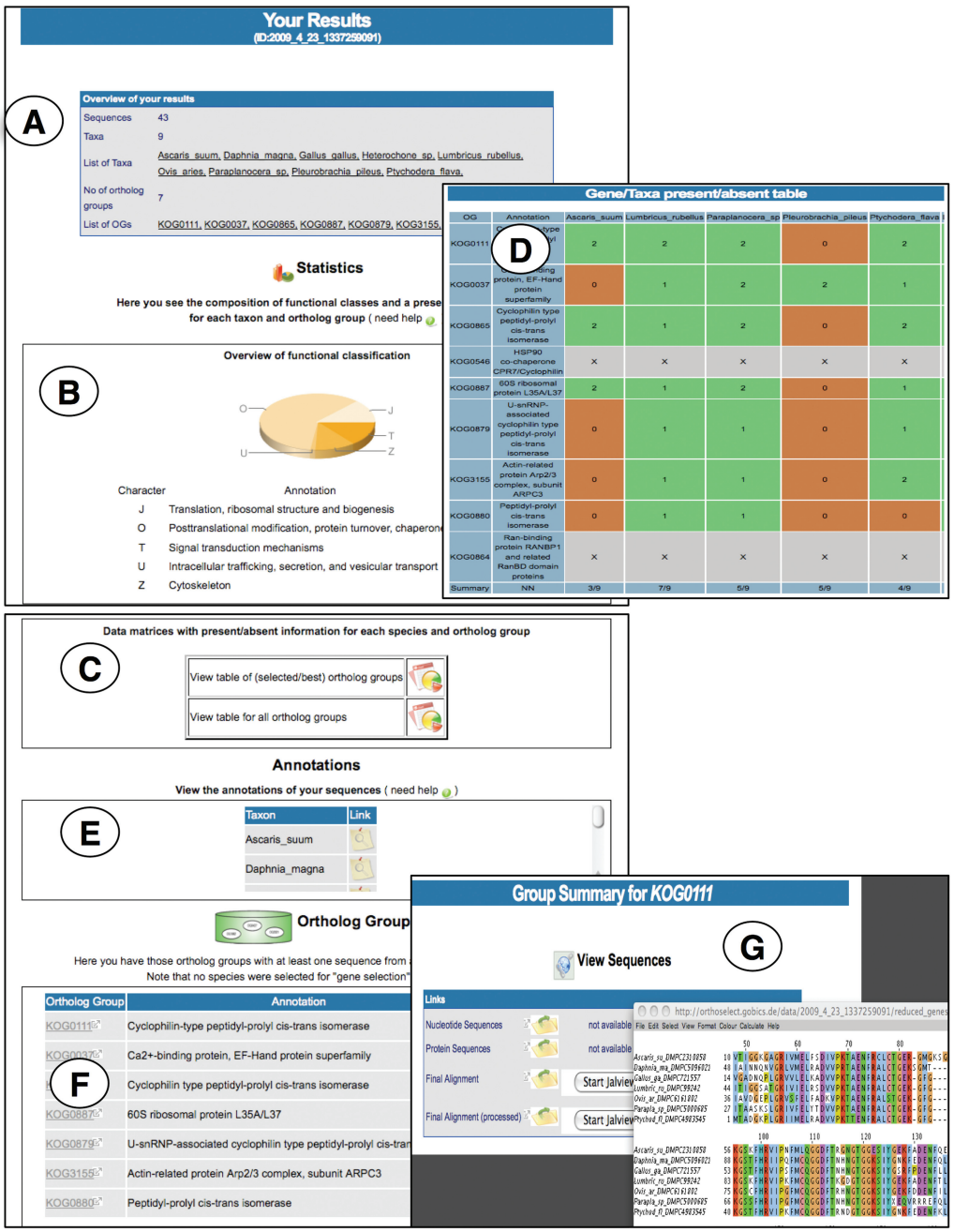

Figure 2. The output of the OrthoSelect web server. Besides a general overview page of the results (**A**), our web server reports information about functional annotations (**B**), a gene/taxa presence/absence matrix (**C**, **D**), annotations for each taxon (**E**), as well as an overview of the orthologous groups (**F**). Additionally, for each orthologous group the resulting alignments are visualized using the Jalview (17) applet (**G**).

OUTPUT

Having generated EST libraries for the species under study, one of the main questions that arise is what genes those EST libraries have in common. These set of common genes can be used as a base for subsequent phylogenetic analysis. Our web server outputs those genes present in all EST libraries, but also provides additional information that will help the user to interpret the data and to decide which data are useful as input for phylogeny programs. The web interface offers a wide range of diagrams, charts, tables, etc. to supply the user with useful information (Figure 2). The most important part is the graphical representation of individual OGs with all assigned and translated EST sequences, and an overview of its taxonomical composition. Single sequences can be viewed along with their translation, as well as the computed multiple sequence alignment prior and subsequent to the final post-processing step in which the program Gblocks or Aliscore are used. The alignment is visualized using the Jalview (17) applet. The web server outputs an overview of

the ESTs' functional classifications and OG assignments as a data matrix with presence/absence information for each gene and species in the study, and annotations for each species. The data matrix shows how many sequences from which taxa have been assigned to an OG. This way, the user can easily select OGs with all or a certain percentage of taxa present.

Besides an overview of all OGs with sequences assigned ('All orthologous groups'), OrthoSelect automatically builds a subset of OGs ('Best orthologous groups') that have either at least three different taxa or the predefined taxa present. The 'all orthologous groups' contain all orthologous groups to which sequences have been assigned, whereas the 'best orthologous groups' only contain one sequence per taxon (see Methods section).

The results page is intended to give the user an elaborate overview and useful information, but also provides all results to be downloaded for further examination and use in phylogenetic studies.

DESIGN AND IMPLEMENTATION

The OrthoSelect server consists of a web interface, a MySQL database management system (DBMS), and the core program OrthoSelect. The web interface for OrthoSelect has been constructed using Grails, which is a web application framework that uses the Groovy scripting language on the Java platform to help standardize the development of web interfaces (http://www.grails.org). Grails follows the idea of keeping data and web pages separate with a controller functioning as a mediator between them. All jobs are split into equal chunks to be computed in parallel on our computer cluster, and all data is stored in the DBMS. The average runtime for 20 000 ESTs with an approx. length of 500 bp is 5 h.

ACKNOWLEDGEMENTS

We thank Katharina Hoff for critically reading the manuscript, Dirk Erpenbeck for testing the web server and Rasmus Steinkamp for supporting the deployment of the web server.

FUNDING

German Research Foundation (DFG, Project Wo896/6-1,2) within DFG Priority Programme SPP 1174 'Deep Metazoan Phylogeny'. Funding for open access charge: Deptartment of Bioinformatics, Georg-August-Universität Göttingen.

Conflict of interest statement. None declared.

REFERENCES

1. Bourlat,S.J., Juliusdottir,T., Lowe,C.J., Freeman,R., Aronowicz,J., Kirschner,M., Lander,E.S., Thorndyke,M., Nakano,H. and Kohn,A.B. (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum xenoturbellida. *Nature*, **444**, 85–88.
2. Dunn,C.W., Hejnol,A., Matus,D.Q., Pang,K., Browne,W.E., Smith,S.A., Seaver,E., Rouse,G.W., Obst,M. and Edgecombe,G.D. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.
3. Delsuc,F., Brinkmann,H., Chourrout,D. and Philippe,H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.
4. Philippe,H., Derelleand,R., Lopez,P., Pick,K., Borchiellini,C., Boury-Esnault,N., Vacelet,J., Renard,E., Houliston,E. and Queinnec,E. (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biol*, **19**, 706–712.
5. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Ann. Rev. Genet.*, **39**, 309–338.
6. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
7. Tatusov,R., Fedorova,N., Jackson,J., Jacobs,A., Kiryutin,B., Koonin,E., Krylov,D., Mazumder,R., Mekhedov,S. and Nikolskaya,A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
8. Zhou,Y. and Landweber,L.F. (2007) BLASTO: a tool for searching orthologous groups. *Nucleic Acids Res.*, **35**, W678–W682.
9. Lottaz,C., Iseli,C., Jongeneel,C.V. and Bucher,P. (2003) Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, **19(Suppl. 2)**, ii103–ii112.
10. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.
11. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
12. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
13. Subramanian,A., Kaufmann,M. and Morgenstern,B. (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.
14. Durbin,R., Eddy,S. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
15. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402,.
17. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.